Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles
Avignon, France, 27 juin au 1er juillet 2022

La Chine de Buffon : édition numérique et exploration sémantique de l'*Histoire naturelle* (1749-1789)

Axel Le Roy, Motasem Alrahabi, Glenn Roe
ObTIC - Sorbonne Université, 75005 Paris, France
prenom.nom@sorbonne-universite.fr

RÉSUMÉ

Nous présentons un travail en cours sur la structuration et l'exploration d'un grand corpus textuel de Georges-Louis de Buffon, célèbre naturaliste français du XVIII[e] siècle. Il s'agit d'éditer en XML-TEI les trente-six volumes de son *Histoire naturelle* et d'effectuer une première exploration autour de la thématique des animaux chinois. Afin de comprendre la représentation du monde chinois et plus particulièrement la construction et la discussion des savoirs sur les animaux dans l'œuvre de Buffon, nous avons commencé à explorer le corpus selon une approche symbolique à base de lexique. Celle-ci permet d'identifier dans les textes les passages porteurs de modalités subjectives: opinions, sentiments ou émotions. Malgré la simplicité de notre approche, les résultats nous ont permis de faire des constats intéressants sur la critique des sources chez Buffon, sur sa description des animaux et sur son observation des pratiques chinoises.

ABSTRACT

Buffon's China: digital editing and semantic exploration of Natural History (1749-1789).
We present work in progress on the structuring and exploration of a large textual corpus containing the works of Georges-Louis de Buffon, the famous 18th-century French naturalist. Our aim is to carry out an initial exploration of the corpus focused on the theme of Chinese animals over the thirty-six volumes of his *Histoire naturelle* encoded in the TEI-XML format. In order to understand the representation of the Chinese world and more particularly the construction and discussion of knowledge about animals in Buffon's work, we began to explore the corpus using a symbolic approach based on pre-existing lexicon. This approach allows us to identify passages in the texts that carry subjective modalities: opinions, feelings or emotions. Despite the simplicity of our approach, the results have allowed us to make some interesting observations about Buffon's source criticism, his description of animals and his observation of Chinese practices.

MOTS-CLÉS : Histoire naturelle, Buffon, Chine, annotation, jugement critique, lecture attentive, corpus, XML-TEI.

KEYWORDS: Natural History, Buffon, China, Annotation, Critical Judgement, Close Reading, Corpus, TEI-XML.

Atelier TAL et Humanités Numériques (TAL-HN), pages 28–35.

1 Introduction

Au tournant des XVII[e] et XVIII[e] siècles, de nombreux textes et objets naturalistes circulaient à l'échelle du monde (Cook, 2017 ; Grasskamp, 2018). Compilés dans les livres, les savoirs sur la nature furent construits à partir des textes d'auteurs anciens ou produits dans des espaces lointains tels que l'Asie. Dans cet article, nous présentons une analyse automatique de textes d'histoire naturelle, et plus particulièrement des ouvrages de Georges-Louis de Buffon, célèbre naturaliste français du XVIII[e] siècle[1]. Nous nous intéressons à la construction des savoirs sur les animaux chinois et nous souhaitons comprendre le regard que portait Buffon sur leur diversité. Quels sont les animaux chinois présentés dans l'*Histoire naturelle* ? Sur quels objets Buffon fait-il porter ses critiques ? Quelles représentations du monde chinois l'*Histoire naturelle* offre-t-elle à ses lecteurs ? On peut aussi se demander si l'*Histoire naturelle* est un lieu d'expression de sinophilie ou de sinophobie, dans un XVIII[e] siècle où les deux visions se succèdent (Étiemble, 1989).

2 Corpus de travail et prétraitement

Afin de constituer un échantillon riche de nombreuses références à l'Empire des Qing, couvrant la période de 1749 à 1789, notre première tâche était de structurer les trente-six volumes de l'*Histoire naturelle* de Buffon en XML-TEI. Cette étape, soutenue par la BnF, et plus particulièrement par Gallica, consistait à réaliser un export XML de la base de données d'origine[2], à reconstruire les livres en rassemblant les pages d'un même volume et en reliant les notes de bas de page aux appels de note. Nous avons ensuite pu enrichir l'édition numérique en complétant les métadonnées, en retravaillant la structure des ouvrages et en ajoutant un encodage spécifique pour les tables des matières et autres index qui se trouvent en début et en fin d'ouvrage.

3 Méthodologie et ressources linguistiques

L'analyse de la subjectivité dans le langage (opinions, sentiments et émotions) est un domaine de recherche très actif en Traitement Automatique des Langues et en Apprentissage Automatique depuis le début une vingtaine d'années (Turney, 2002 ; Wiebe et al., 2005 ; Pang & Lee, 2008 ; Balahur et al., 2011 ; Birjali et al., 2021).

---

[1] Voir le travail de Daniel Mornet sur la présence des œuvres de Buffon dans des catalogues de privées (1750-1781) (Mornet, 1910).

[2] Édition électronique de l'*Histoire Naturelle* de Buffon, Pietro Corsi, Thierry Hoquet et Stéphane Pouyllau, Paris, Centre Alexandre-Koyré, CRHST/CNRS, 2005. Actuellement, le site www.buffon.cnrs.fr n'est pas publiquement accessible.

Dans le cadre de notre étude, afin d'analyser les jugements critiques concernant les animaux chinois dans l'œuvre de Buffon, nous nous sommes basés sur une approche symbolique. Les systèmes d'apprentissage automatique nécessitent en effet des corpus préannotés dont nous ne disposons pas pour le XVIII[e] siècle, et plus particulièrement pour ce genre textuel. Notre objectif final est de mettre à disposition de la communauté de telles données enrichies d'étiquettes sémantiques fines en lien avec les modalités subjectives. Dans ce sens, nous avons entrepris une annotation automatique du corpus à l'aide de dictionnaires de motifs linguistiques préalablement créés et classés dans des catégories telles que les opinions, les prises de position, l'appréciation esthétique, la dépréciation éthique, etc. Les motifs permettent de détecter uniquement les modalités, sans les arguments sous-jacents, c'est à dire la source ou la cible du jugement[3]. Ils sont composés de marqueurs saillants (verbes, adjectifs, adverbes, etc.) et classés selon leur proximité sémantique et leur polarité positive ou négative[4]. Voici par exemple quelques catégories à polarité négative : désaccord (p. ex. *contester, adversaire, contre le gré*…), dépréciation (p. ex. *maladroit, déplaire, inutilement*…), indignation (p. ex. *désapprouver, révoltant, indignation*…), colère (p. ex. *furieux, à bout de nerfs, s'énerver*…), souffrance (p. ex. *au supplice, endurer, misérable*…).

Ces ressources ont été initialement développées sur des corpus modernes comme les textes journalistiques (Alrahabi, 2016) et les textes de critique littéraire du XIX[e] siècle (Riguet & Alrahabi, 2018). Cette expérience soulève donc la question de la variation diachronique des marqueurs linguistiques et leur adéquation au domaine d'application.

L'annotation automatique se fait au niveau de la surface, sans analyse morphosyntaxique, à l'aide d'un outil développé par notre équipe qui permet au préalable de segmenter les paragraphes en phrases. Voici un exemple de phrase annotée avec deux catégories de prise de position telles que « accord » et « désaccord » :

« *Au reste, en refusant à l'éperonnier le nom de paon de la Chine, je ne fais que me conformer aux témoignages des Voyageurs qui assurent que dans ce vaste pays, on ne voit de paons que ceux qu'on y apporte des autres contrées* ». (*Histoire naturelle des oiseaux*, vol. 2, 1771, p. 371).

Nous avons enrichi la base des motifs avec des marqueurs d'intensité (p. ex. *essentiellement, excessif, positivement*, etc.) et des marqueurs de négation (p. ex. *ne pas, jamais, aucun*, etc.).

Une fois annotés, les textes XML-TEI sont ensuite indexés dans Ariane (Alrahabi, 2021)[5], une interface d'exploration et de fouille de textes (Figure 1). Celle-ci permet à l'utilisateur d'effectuer des recherches par mots-clés sur l'ensemble du corpus, et de croiser les termes identifiés avec les phrases annotées. Plusieurs paramètres de filtrage par métadonnées, des statistiques et des visualisations sont

[3] Il serait intéressant d'introduire ultérieurement de nouvelles analyses afin d'identifier ces éléments (voir par exemple les travaux sur la Stance Detection, Hardalov et al. 2021).

[4] La base contient environ 70 catégories et plus de 3000 motifs linguistiques. Nous avons écarté de cette étude les catégories neutres comme l'observation, l'assertion, la définition, etc.

[5] Voir https://obvil.huma-num.fr/ariane/

proposés afin de simplifier le passage d'une approche quantitative à une lecture attentive (*close reading*).

FIGURE 1: Interface de lecture dans Ariane

Afin de faire ressortir les passages qui nous intéressent – à savoir ceux qui concernent les animaux présents en Chine –, un premier filtrage par métadonnées dans l'interface nous permet de limiter le corpus aux quinze volumes sur les quadrupèdes[6], aux neuf volumes sur les oiseaux[7] et aux trois volumes de suppléments qui portent sur les animaux : soit un peu plus de 4 millions de mots. Ensuite, à partir de la requête « Chine » et des mots apparentés (*chinois, Pékin, Macao*, etc.), le système nous renvoie 258 phrases dans 18 livres, dont 119 sont porteuses de jugements critiques.

4 Résultats et interprétations

Le fait de passer d'une grande masse de données à une première lecture attentive des segments annotés permet de mettre en lumière trois formes de subjectivités présentes sous forme de jugement dans ce

6 *Histoire naturelle des quadrupèdes*.
7 *Histoire naturelle des oiseaux*.

corpus d'ouvrages scientifiques[8] : la critique des sources, la description des animaux, l'observation des pratiques chinoises.

Le premier ensemble de jugements est apparenté au débat intellectuel, à la critique des sources, caractéristiques de la construction sociale des savoirs. Le système repère plus exactement des jugements dépréciatifs portés dans l'*Histoire naturelle* sur les sources qui y sont mobilisées, notamment concernant des questions d'identification et de classification des animaux : deux tâches centrales dans le travail naturaliste, sources de nombreux débats, surtout pour les espèces nouvellement découvertes. Exemple:

« La Relation Hollandoise qui a pour titre, l'Ambassade de la Chine, fait une description de cet animal [le rhinocéros] tout-à-fait fausse[9] » (*Histoire naturelle*, vol. 11, 1754, p. 194) [étiquette "Incorrect"].

Mais les résultats obtenus sur ce corpus dépassent la seule question de la critique des sources citées. L'outil met en effet en exergue un deuxième ensemble de jugements : ceux portés par l'auteur sur les animaux. Exemple :

« Le tricolor huppé [...], est ce beau faisan dont on dit que les plumes se vendent à la Chine plus cher que l'oiseau même » (*Histoire naturelle des oiseaux*, vol. 2, 1771, p. 359) [étiquette "Appréciation-Esthétique"].

Les jugements portés sur le comportement, l'anatomie et la beauté – identifiée par les catégories Appréciation ou de Dépréciation esthétique – sont constitutifs des représentations littéraires de la nature chinoise diffusées par l'*Histoire naturelle*.

Nous observons enfin un troisième ensemble de jugements, portés sur les pratiques chinoises d'élevage, de domestication et de chasse.

« Le han-ta-han (disent les Missionnaires) est un animal qui ressemble à l'élan ; la chasse en est commune dans le pays des Solons, et l'Empereur Kam-hi prenoit quelquefois plaisir à cet amusement » (*Histoire naturelle*, vol. 12, 1769, p. 90) [étiquettes "Appréciation" et "Plaisir"].

De telles descriptions de relations entre les humains et les animaux s'inscrivent dans la continuité de celles diffusées – par – dans des récits de voyages qui ont parfois été mobilisés par Buffon.

---

[8] Nous considérons ainsi le texte naturaliste comme une construction littéraire (Roger, 1963) – sans négliger qu'il est aussi le produit d'un contexte social (Shapin et Schaffer, 1993) – dans le sillage du travail de Nathalie Vuillemin (Vuillemin 2009).

[9] Cette critique d'une description du rhinocéros publiée exactement un siècle plus tôt à Amsterdam à partir des notes accumulées par Johan Nieuhof – membre de l'ambassade hollandaise menée en Chine entre 1655 et 1657 – est un exemple de la circulation des textes.

5 Évaluation

Notre base de motifs linguistiques contient plus de soixante-dix catégories sémantiques. Vu la difficulté de la tâche d'évaluation, nous avons procédé dans un premier temps à l'analyse de la qualité des annotations à partir d'une sous-partie du corpus. Nous avons exporté d'Ariane les passages qui contiennent les termes *Chine, chinois, Pékin et Macao*. Ensuite, nous avons demandé à deux chercheurs d'annoter l'ensemble de ces extraits, en attribuant à chaque phrase 0 ou n étiquettes à partir de notre base de catégories sémantiques.

L'accord global entre annotateurs selon la mesure de Kappa est d'une concordance importante : 0,63. Les divergences concernent souvent les nuances entre des catégories fines, par exemple entre appréciation esthétique, appréciation éthique, appréciation psychoaffective et appréciation intellectuelle.

En comparant l'ensemble des annotations automatiques aux annotations manuelles[10], nous obtenons un score global de 92 % pour la précision et 81 % pour le rappel. La F-mesure s'élève à 86 %. Des erreurs d'annotation automatique ont été observées. Elles sont liées à la polysémie de certains marqueurs homographes : *élan, bécasse, sauvage, océan Pacifique*, fleuve *Amour*, *fécond, grogner*, etc. Employés dans le premier ensemble de critique des sources, ces marqueurs ont un sens connoté, mais quand ils concernent la description des animaux et des pratiques chinoises, ils pourraient générer du bruit. Le changement du genre textuel a donc un impact sur la qualité des annotations, ce qui nécessite l'adaptation de certains marqueurs. D'autres marqueurs peuvent prendre dans notre corpus un sens différent mais pas opposé. Le verbe "*Tuer*" par exemple n'est pas systématiquement lié à la catégorie "Violence", mais aussi aux traditions d'élevage :

« On tua un de ces animaux chez M. le duc de Richemont, mais la chair ne s'en est pas trouvée si bonne que celle du bœuf. » (*Supplément à l'Histoire naturelle*, vol. 3, 1776, p. 63).

Les cas faux négatifs sont principalement dus à la couverture de la base des motifs, aux traces qui subsistent de l'ancienne orthographe du corpus (p. ex. *obéissan* → *obéissant*) et aux erreurs d'océrisation (p. ex. *efpèce* → *espèce*), principalement dans la transcription du sixième volume des *Suppléments*.

6 Conclusion et perspectives

Cette première exploration du corpus nous a permis de faire des constats intéressants sur la critique des sources chez Buffon, sur sa description des animaux et sur son observation des pratiques chinoises. En plus du travail de structuration en XML-TEI du corpus de Buffon, l'apport de ce travail réside notamment dans l'approche de fouille utilisée qui se veut opérationnelle et qui offre un moyen

[10] Nous avons pris en compte uniquement les annotations en commun entre les deux juges.

rapide pour l'analyse de grands corpus textuels en passant du distant au *close reading*, l'analyse se faisant à l'échelle du corpus tout entier comme à l'échelle de la phrase annotée. En ce qui concerne le processus d'annotation, la détection des marqueurs d'intensité et de négation doit être améliorée. Nous envisageons également de poursuivre la modernisation de l'orthographe de ce corpus du XVIII^{e} siècle et corriger les erreurs de transcription qui subsistent avant d'annoter un grand corpus de référence pour l'entraînement d'un modèle d'apprentissage automatique.

Si les jugements portés par Buffon sont parfois très polarisés dans le détail (sur les pratiques chinoises, la beauté des oiseaux chinois) ils ne portent pas tous sur le monde chinois (critique des auteurs) et, à l'échelle de l'œuvre, aucune position sinophile ou sinophobe globale et tranchée ne se distingue. Dans l'ensemble, nous avons 66 annotations positives et 53 négatives dans des passages sur la Chine. L'*Histoire naturelle* semble occuper sur cette question une position médiane, qui reflète la position chronologiquement centrale de l'œuvre, entre l'enthousiasme sinophile des premières décennies du XVIII^{e} siècle et la sinophobie croissante au tournant des XVIII^{e} et XIX^{e} siècles.

Une poursuite intéressante du travail consiste à élargir la focale en s'intéressant aux textes sur les animaux d'autres espaces géographiques tels que le Japon, le Siam, le Brésil ou encore le Pérou, puis à comparer les résultats entre ces différents ensembles. Le corpus pourrait également être élargi à d'autres ouvrages sur les animaux comme l'*Histoire naturelle* des poissons de Cuvier et Valenciennes, ainsi qu'à d'autres domaines de l'histoire naturelle comme la botanique ou les volumes sur la minéralogie de Buffon que nous avons écartés dans cette étude.

Références

ALRAHABI M. (2010). *EXCOM-2: plateforme d'annotation automatique de catégories sémantiques. Applications à la catégorisation des citations en français et en arabe.* Thèse de doctorat, Université Paris-Sorbonne.

ALRAHABI M. (2016). E-Quotes : un outil de navigation textuelle guidée par les annotations sémantiques. In *Actes de TALN 2016*, Paris. ATALA.

ALRAHABI M. (2021). Ariane: dispositif de fouille et de lecture synthétique de textes. In *Actes de DigitAl Humanities and cuLtural herItAge: data and knowledge management and analysis (Atelier Dahlia)*, Jan 2021, Montpellier. Hal : hal-03167271.

BALAHUR A., HERMIDA J., MONTOYO A. & MUÑOZ R. (2011). EmotiNet: A Knowledge Base for Emotion Detection in Text Built on the Appraisal Theories. In *Natural Language Processing and Information Systems*, volume 6716, p. 27-39. Doi : 10.1007/978-3-642-22327-3_4.

COOK H. J. (2007). *Matters of exchange: commerce, medicine, and science in the Dutch Golden Age*, New Haven.

ÉTIEMBLE R. (1989). *L'Europe chinoise. II. De la sinophilie à la sinophobie*, Gallimard.

GRASSKAMP A. & JUNEJA M. (2018). *EurAsian matters: China, Europe, and the transcultural object, 1600-1800*, Springer.

HARDALOV M., ARORA A., NAKOV P. & AUGENSTEIN I. (2021). A Survey on Stance Detection for Mis- and Disinformation Identification. In *ArXiv*. Doi : 10.48550/arXiv.2103.00242

LILTI A. (2007). Querelles et controverses: Les formes du désaccord intellectuel à l'époque moderne. In *Mil neuf cent. Revue d'histoire intellectuelle*, volume 25, p. 13-28. Doi : 10.3917/mnc.025.0013.

HOQUET T. (2007). *Buffon/Linné. Éternels rivaux de la biologie ?*, Dunod.

MAROUANE B., KASRI M. & BENI-HSSANE A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends Knowl. In *Knowledge-Based Systems*, volume 226. Doi : 10.1016/j.knosys.2021.107134.

MCMAHON M. D. (2001), *Enemies of the enlightenment : The French counter-enlightenment and the making of modernity*, Oxford University Press.

MORNET D. (1910). Les enseignements des bibliothèques privées (1750-1781). In *Revue d'histoire littéraire de la France* 17, p. 449-496.

PANG B. & LEE L. (2008). Opinion Mining and Sentiment Analysis. In *Foundations and Trends in Information Retrieval*, vol. 2, no 1-2, p. 1-135. Doi : 10.1561/1500000001.

PARADIS S. (2008). *Imagination, jugement, génie : la fabrique des quadrupèdes dans l'Histoire naturelle de Buffon (1707-1788)*. Thèse de doctorat, Université de Laval.

RIGUET M. & ALRAHABI M. (2018). Pour une analyse automatique du Jugement Critique: les citations modalisées dans le discours littéraire du XIXe siècle. In *DHQ: Digital Humanities Quarterly* 1.12.

ROGER J. (1963). *Les sciences de la vie dans la pensée française du XVIIIe siècle. La génération des animaux de Descartes à l'Encyclopédie*, Albin Michel.

ROMAN H. (2018). *The Language of nature in Buffon's Histoire naturelle*, Oxford University Studies in the Enlightenment.

SHAPIN S. & SCHAFFER S. (1993). *Léviathan et la pompe à air, Hobbes et Boyle entre science et politique*, La Découverte.

TURNEY P. (2002). Thumps up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Actes de The 40th Annual Meeting of the Association for Computational Linguistics*.

VUILLEMIN N. (2009). *Les beautés de la nature à l'épreuve de l'analyse*, Presses Sorbonne Nouvelle.

WIEBE J., WILSON T. & CARDIE C. (2005). Annotating Expressions of Opinions and Emotions in Language. In *Language Resources and Evaluation*, volume 39, issue 2-3, p. 165-210.